National Human Genome Research Institute

Semantic Wiki for Collaborative Annotation of Chemical Biology Data

Ajay[1], A. Zavaleta[2], A. Frazin[2], S. Chintala[2], and H. Cheung[2]

[1]Cheminformatics Research Centers, National Human Genome Research Institute
[2]Division of Computational Bioscience, Center for Information Technology

CIT Center for Information Technology

Abstract

As a central component of the Molecular Libraries Program (MLP), a NIH Roadmap project, the Molecular Libraries Program Center Network (MLPCN) is an effort to build high-throughput based, chemo-biological screening to test and understand the interaction between small molecular chemical structures and biological targets.

Resources for exchanging data and information have been ubiquitous in biology. The effort to enable cross-querying across databases with different types of data has led to extensive work in creating controlled-vocabularies and ontologies in multiple areas of biology. Chemical biology data poses specific challenges that require scientists from different disciplines to collaborate. Semantic wiki technology helps to merge the two strains of ontologies and scientific annotations. As a collaborative web-authoring tool, a semantic wiki makes it easier for scientists to contribute and collaborate.

PubChem is the official data repository for the MLP. The semantic wiki that is being developed as a cheminformatics tool would extract fundamental parameters of the screening data from PubChem into existing pre-defined controlled vocabularies and allow scientists to further annotate both biological, as well as chemical data. This would make high-throughput screening data more useful for gaining insight from the chemo-biological processes and would accelerate the development of new therapeutics.

Problem Statement

Each probe project undertaken by the MLPCN goes through multiple biological assays each of which are exposed to a number of small molecules. Each biological assay can be one of three types:

1. Isolated molecular target assays
2. Cell-free multi-component assays
3. Cell or even organism based assays

The formats (e.g., plate type), controls, output, detector (e.g., imaging microscope) etc. can vary widely. In the same fashion, the sources of the compounds used in the assays and their rationale can vary widely.

To ease the task of interpretation of the results we hope to build a set of **"minimum information for chemical biology experiments**" when the results are reported out. This implies building our own set of controlled vocabularies as well as re-using existing ones. Additionally, annotating the experiments with a sufficiently dense set of controlled meta-information like ATCC number for identifying cell-lines, Uniprot Id for the protein, the GO biological process being studied will help in generating multiple ways to browse the set of experimental results.

Problem Solution

To address these problems we are adopting a Semantic Wiki for the MLPCN scientists to use for

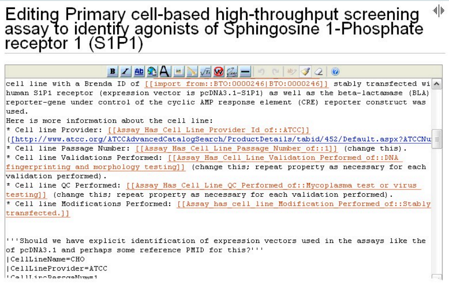
1. Generating appropriate controlled vocabularies
2. Annotating each experiment with suitable meta-information
3. Tools for navigating the entire set of experimental results from multiple perspectives

Generating Appropriate Controlled Vocabularies

Semantic MediaWiki (SMW), an enhancement of the MediaWiki software, provides an open content development platform that would allows the generation of appropriate controlled vocabularies. It also provides a mechanism for collaboration between multiple stakeholders for vocabularies review, annotation, and recommendation.

The infrastructure of the SMW provides the following benefits:

- Enables property, inline queries, and new ways to explore content
- Creates a collaborative environment for collective editing and harvesting collective knowledge.
- Keeps a complete record of change history
- Combines structured and unstructured knowledge for reusable contents
- Provides a formal RDF export for inferencing



Editing Primary cell-based high-throughput screening assay to identify agonists of Sphingosine 1-Phosphate receptor 1 (S1P1)

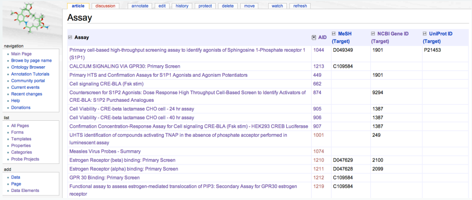
```
cell line with a Brenda ID of [[import from::BTO:0000246|BTO:0000246]] stably transfected wi
human S1P1 receptor (expression vector is pcDNA3.1-S1P1) as well as the beta-lactamase (BLA)
reporter-gene under control of the cyclic AMP response element (CRE) reporter construct was
used.
Here is more information about the cell line:
* Cell line Provider: [[Assay Has Cell Line Provider Id of::ATCC]]
([http://www.atcc.org/ATCCAdvancedCatalogSearch/ProductDetails/tabid/452/Default.aspx?ATCCNu
* Cell line Passage Number: [[Assay Has Cell Line Passage Number of::1]] (change this).
* Cell line Validations Performed: [[Assay Has Cell Line Validation Performed of::DNA
fingerprinting and morphology testing]] (change this; repeat property as necessary for each
validation performed).
* Cell line QC Performed: [[Assay Has Cell Line QC Performed of::Mycoplasma test or virus
testing]] (change this; repeat property as necessary for each validation performed).
* Cell line Modifications Performed: [[Assay has cell line Modification Performed of::Stably
transfected.]]


'''Should we have explicit identification of expression vectors used in the assays like the
of pcDNA3.1 and perhaps some reference PMID for this?'''
|CellLineName=CHO
|CellLineProvider=ATCC
```

Figure 1: Assays Annotation.

Annotating Experiments with Meta-Information

PubChem contains thousands of chemical, biological, and bioassay screening data. The BioAssayOnDemand extension assist scientists to extract and aggregate relevant chemical biology data from PubChem for the generation of meta-data and semantic authoring and annotations. BioAssayOnDemand is a module for the PagesOnDemand extension. It generates wiki articles about bioassays in the PubChem database. Its greatest utility is in adding semantic content to the extracted data.

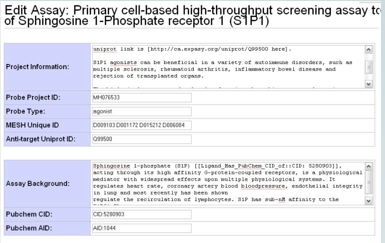


Assay

Assay	AID	MeSH (Target)	NCBI Gene ID (Target)	UniProt ID (Target)
Primary cell-based high-throughput screening assay to identify agonists of Sphingosine 1-Phosphate receptor 1 (S1P1)	1044	D049349	1901	P21453
CALCIUM SIGNALING VIA GPR30: Primary Screen	1213	C109584		
Primary HTS and Confirmation Assays for S1P1 Agonists and Agonism Potentiators	449		1901	
Cell signaling CRE-BLA (Fsk stim)	662			
Counterscreen for S1P2 Agonists: Dose Response High Throughput Cell-Based Screen to Identify Activators of CRE-BLA: S1P2 Purchased Analogues	874		9294	
Cell Viability - CRE-beta lactamase CHO cell - 24 hr assay	905		1387	
Cell Viability - CRE-beta lactamase CHO cell - 40 hr assay	906		1387	
Confirmation Concentration-Response Assay for Cell signaling CRE-BLA (Fsk stim) - HEK293 CREB Luciferase	907		1387	
UHTS identification of compounds activating TNAP in the absence of phosphate acceptor performed in luminescent assay	1001		249	
Measles Virus Probes - Summary	1074			
Estrogen Receptor (beta) binding: Primary Screen	1210	D047629	2100	
Estrogen Receptor (alpha) binding: Primary Screen	1211	D047628	2099	
GPR 30 Binding: Primary Screen	1212	C109584		
Functional assay to assess estrogen-mediated translocation of PIP3: Secondary Assay for GPR30 estrogen receptor	1219	C109584		

Figure 2: Assays extracted from PubChem for annotation.

Semantic Forms is an extension to MediaWiki that allows users to create structured forms for adding and editing data. Forms can be created on-the-fly based on existing data, the form definition and the templates that the form outputs



Edit Assay: Primary cell-based high-throughput screening assay to of Sphingosine 1-Phosphate receptor 1 (S1P1)

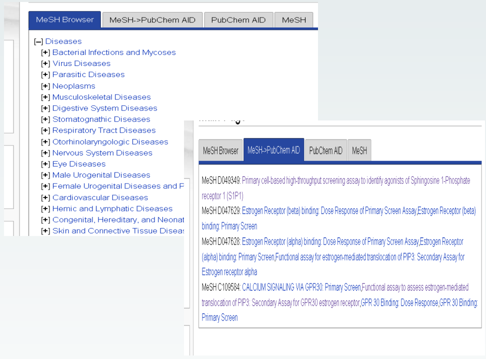
Project Information:	uniprot link is [http://ca.expasy.org/uniprot/Q99500 here]. S1P1 agonists can be beneficial in a variety of autoimmune disorders, such as multiple sclerosis, rheumatoid arthritis, inflammatory bowel disease and rejection of transplanted organs.
Probe Project ID:	MH076533
Probe Type:	agonist
MESH Unique ID	D009103 D001172 D015212 D006084
Anti-target Uniprot ID:	Q99500
Assay Background:	Sphingosine 1-phosphate (S1P) [[Ligand_Has_PubChem_CID_of::CID: 5280903]], acting through its high affinity G-protein-coupled receptors, is a physiological mediator with widespread effects upon multiple physiological systems. It regulates heart rate, coronary artery blood bloodpressure, endothelial integrity in lung and most recently has been shown regulate the recirculation of lymphocytes. S1P has sub-nM affinity to the
Pubchem CID:	CID:5280903
Pubchem AID:	AID:1044

Figure 3: Assays Annotation with Template and Semantic Forms.

Tools for Navigation

The Medial Subject Headings (MeSH) provides an invaluable resource to locate descriptors of possible interest and to display their hierarchical relationship. However, it does not directly connect to database retrieval systems. In order to link MeSH to other sources of biological information, a Semantic MeSH Tree Browser is implemented.

The Semantic MeSH Tree browser is similar to the MeSH Browser, except that it displays relationships between MeSH terms and assay pages in the wiki. This is possible because the semantic annotation of assays allows them to be dynamically queried via properties. For example, below we display the Semantic Tree Browser where each node displays the relationship between a specific MeSH term and Assay IDs.



MeSH Browser
MeSH->PubChem AID
PubChem AID
MeSH
[–] Diseases
[+] Bacterial Infections and Mycoses
[+] Virus Diseases
[+] Parasitic Diseases
[+] Neoplasms
[+] Musculoskeletal Diseases
[+] Digestive System Diseases
[+] Stomatognathic Diseases
[+] Respiratory Tract Diseases
[+] Otorhinolaryngologic Diseases
[+] Nervous System Diseases
[+] Eye Diseases
[+] Male Urogenital Diseases
[+] Female Urogenital Diseases and F
[+] Cardiovascular Diseases
[+] Hemic and Lymphatic Diseases
[+] Congenital, Hereditary, and Neonat
[+] Skin and Connective Tissue Diseas
MeSH Browser
MeSH->PubChem AID
PubChem AID
MeSH

MeSH D049349: Primary cell-based high-throughput screening assay to identify agonists of Sphingosine 1-Phosphate receptor 1 (S1P1)

MeSH D047629: Estrogen Receptor (beta) binding: Dose Response of Primary Screen Assay,Estrogen Receptor (beta) binding: Primary Screen

MeSH D047628: Estrogen Receptor (alpha) binding: Dose Response of Primary Screen Assay,Estrogen Receptor (alpha) binding: Primary Screen,Functional assay for estrogen-mediated translocation of PIP3: Secondary Assay for Estrogen receptor alpha

MeSH C109584: CALCIUM SIGNALING VIA GPR30: Primary Screen,Functional assay to assess estrogen-mediated translocation of PIP3: Secondary Assay for GPR30 estrogen receptor,GPR 30 Binding: Dose Response,GPR 30 Binding: Primary Screen

Figure 4: The Semantic MeSH Tree Browser.

References

[1] Semantic wiki: http://en.wikipedia.org/wiki/Semantic_wiki
[2] PageOnDemand: http://www.mediawiki.org/wiki/Extension:PagesOnDemand
[3] Semantic Forms: http://www.mediawiki.org/wiki/Extension:Semantic_Forms
[4] Category Tree: http://www.mediawiki.org/wiki/Extension:CategoryTree